ASBMT
American Society for Blood and Marrow Transplantation

Oral Chronic Graft-versus-Host Disease Scoring Using the NIH Consensus Criteria

Nathaniel S. Treister,[1,2] Kristen Stevenson, MS,[3] Haesook Kim,[3] Sook-Bin Woo,[1,2] Robert Soiffer,[4,5] Corey Cutler[4,5]

The National Institutes of Health (NIH) Oral chronic Graft-versus-Host Disease (cGVHD) Activity Assessment Instrument is intended to be simple to use and to provide a reproducible objective measure of disease activity over time. The objective of this study was to assess inter- and intraobserver variability in the component and composite scores in patients evaluated with oral cGVHD. Twenty-four clinicians (bone marrow transplant [BMT] oncologists: BMTE, n = 16; BMT midlevel providers: BMT MLP; n = 4; and oral medicine experts [OME], n = 4), from 6 major transplant centers scored high-quality intraoral photographs of 12 patients. The same photographs were evaluated 1 week later by the same evaluators. An intraclass correlation coefficient (ICC) was used to calculate intrarater reliability and interrater agreement was analyzed using a weighted κ statistic: $0 \leq \kappa \leq 0.20$ = poor, $0.21 \leq \kappa \leq 0.40$ = fair, $0.41 \leq \kappa \leq 0.60$ = moderate, $0.61 \leq \kappa \leq 0.80$ = good, $0.81 \leq \kappa \leq 1.00$ = very good. Data on participant experiences and demographics were also collected. Mean interrater reliability for each element was poor to moderate (range: 0.15-0.46). Overall mean kappa scores were highest for ulcers (0.46), followed by erythema (0.23), and lowest for lichenoid (0.15) and mucoceles (0.14). Kappa scores were higher in OME compared with BMTE and BMT MLP in ulcers and erythema (eg, 0.85, 0.44, 0.33 for ulcers, respectively), but similar in lichenoid and mucoceles. Overall intrarater reliability in all groups was very good ($\geq$0.90) and highest for ulcers (0.97, 0.85, 0.94). Although 75% of OME were *comfortable* with their abilities to score the cases, approximately 50% of BMTE and BMT MLP were *uncomfortable*. The majority felt that their evaluations were *accurate*; however, 84% agreed that *formal training is required*. Interrater variability of the oral cGVHD instrument is unacceptable for the purposes of clinical trials. Greater concordance among OME, high intrarater reliability, and participant feedback suggests that formal training may significantly decrease variability. Parallel investigations must be completed using the other organ specific instruments prior to any revision and widespread prospective utilization of these tools as research endpoints.

Biol Blood Marrow Transplant 16: 108-114 (2010) © 2010 American Society for Blood and Marrow Transplantation

KEY WORDS: Oral, Chronic graft-versus-host disease, Hematopoietic cell transplantation

INTRODUCTION

Chronic graft-versus-host disease (cGVHD) is a serious and potentially life-threatening complication of allogeneic hematopoietic stem cell transplantation (allo-HCT) [1-5]. With many hematologic diseases, malignancies, and bone marrow failure syndromes increasingly being treated with alloHCT, and with continually improving supportive care measures and survival rates, cGVHD has become the leading long-term cause of morbidity and mortality after transplantations [1,6-9]. Treatment of cGVHD is often only partially effective, and aside from first-line therapy with high dose corticosteroids, there is no consensus as to what constitutes standard second-line therapy, demonstrating the critical need for large-scale, multi-institutional clinical trials.

One of the main barriers to the conduct of effective clinical research in cGVHD has been the absence of standardized criteria for diagnosis, staging, and

From the [1]Division of Oral Medicine and Dentistry, Brigham and Women's Hospital, Boston, Massachusetts; [2]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, Massachusetts; [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts; [4]Division of Hematologic Malignancies, Dana-Farber Cancer Institute, Boston, Massachusetts; and [5]Department of Medicine, Harvard Medical School, Boston, Massachusetts.

Financial disclosure: See Acknowledgments on page 113.

Correspondence and reprint requests: Nathaniel S. Treister, DMD DMSc, Division of Oral Medicine and Dentistry, Brigham and Women's Hospital, 1620 Tremont Street, 3rd floor, BC-3-028, Boston, MA 02120 (e-mail: ntreister@partners.org).

Received June 22, 2009; accepted September 14, 2009

1083-8791/10/161-00013$36.00/0
doi:10.1016/j.bbmt.2009.09.010

response to therapy [10]. This was recently addressed by the National Institutes of Health (NIH)-sponsored Chronic GVHD Consensus Project, from which a number of consensus documents were published, including response criteria guidelines intended to measure clinical changes over time [11]. Importantly, the cGVHD Activity Assessment instruments for response measurement were designed to be easy to use not by organ-specific specialists (eg, dermatologists and oral medicine specialists), but rather by transplant physicians, nurses, and physician assistants. None of these instruments have been prospectively validated, and the extent to which they may truly be effective in discerning clinically meaningful changes remains unclear.

The oral cavity is one of the most frequently affected target organs of cGVHD, with prominent mucosal changes including *erythema* (redness), *lichenoid hyperkeratotic* changes (white reticulation and/or plaques), *ulceration* (yellow-to-white pseudomembranes), and *mucoceles* (mucous-filled vesicles) that are well-visualized during routine clinical examination [12-14]. These features make the mouth an excellent representative site for prospective evaluation of the organ-specific response criteria. The purpose of this study was to evaluate inter- and intraobserver variability in the Oral cGVHD Activity Assessment instrument component and composite scores, and to identify any significant impediments to its use.

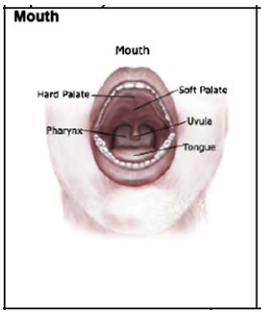
METHODS

Twenty clinicians from 6 institutions, including oncologists (denoted BMTE) and non-MD clinical staff (nurses, physician assistants, and dental hygienists; denoted BMT MLP) with experience in evaluating patients with cGVHD reviewed high-quality intraoral photographs of patients with oral cGVHD. Photographs were obtained as part of routine clinical evaluation at the Center for Oral Disease, Brigham and Women's Hospital, Boston, MA. All patients consented to the use of their photographs for teaching and research purposes. Twelve cases representing a spectrum of clinically active and resolved cGVHD, ranging from mild changes to extensive ulcerations, were selected for the study and printed in series on photographic paper. To establish a standard to compare the evaluators' responses, 4 oral medicine specialists (denoted OME) also evaluated the cases via the same protocol, for a total of 24 raters. This study was approved by the Dana-Farber/Harvard Cancer Center's Office for Human Research Subjects.

Oral cGVHD scoring

The NIH oral cGVHD Activity Assessment is a 0-15 point objective clinical scoring system that takes into account (1) severity and extent of *erythema*, (2) extent of *lichenoid hyperkeratotic changes*, (3) extent of *ulcerations*, and (4) the number of *mucoceles* (Figure 1). Evaluators were instructed to score the cases based on a global assessment of clinical signs rather than on any given anatomic site, according to published guidelines (http://www.asbmt.org) [15]. For example, to assign a score of "moderate" ulceration, defined as ≤20% involvement, 20% of all evaluated sites would need to demonstrate ulceration, not 20% of a single anatomic site, such as the right buccal mucosa. Evaluators were instructed to consider any sites that were

Figure 1. NIH Oral cGVHD Activity Assessment scoring instrument.



Mucosal change	No evidence of cGVHD		Mild		Moderate		Severe	
Erythema	None	0	Mild erythema or moderate erythema (<25%)	1	Moderate (≥25%) or Severe erythema (<25%)	2	Severe erythema (≥25%)	3
Lichenoid	None	0	Hyperkeratotic changes(<25%)	1	Hyperkeratotic changes(25-50%)	2	Hyperkeratotic changes (>50%)	3
Ulcers	None	0	None	0	Ulcers involving (≤20%)	3	Severe ulcerations (>20%)	6
Mucoceles*	None	0	1-5 mucoceles	1	6-10 scattered mucoceles	2	Over 10 mucoceles	3
			*Mucoceles scored for lower labial and soft palate only				Total score for all mucosal changes	

Figure 2. Representative series of oral cGVHD photographs.

not included in the series of photographs as "normal," or unaffected.

Identical study packets were distributed by mail to each participant that included detailed instructions, clinical photographs, an evaluator questionnaire, scoring forms, and a postage-paid return envelope. Evaluations of the same 12 cases were completed independently 2 times, 1 week apart; evaluators were explicitly instructed not to review their first scores while completing the second scoring sheet. After the second evaluation, raters answered a series of questions (using a 5-point scale) about their ease and comfort with scoring the cases, and entered comments to note any areas that seemed particularly problematic or unclear. Completed study materials were returned in the provided postage paid return envelope (Figure 2).

Statistical Analyses

Agreement between raters (interrater agreement) was analyzed using a weighted Kappa (κ) statistic, which compensates for equivalent ratings because of chance [16]. The degree of agreement was interpreted as follows: $0 \leq \kappa \leq 0.20$ = *poor*, $0.21 \leq \kappa \leq 0.40$ = *fair*, $0.41 \leq \kappa \leq 0.60$ = *moderate*, $0.61 \leq \kappa \leq 0.80$ = *good*, and $0.81 \leq \kappa \leq 1.00$ = *very good* [17]. To investigate the reliability of the instrument and intrarater agreement, cases were evaluated by raters independently on 2 occasions, 1 week apart. Intrarater reliability was estimated using Pearson's correlation coefficient and an intraclass correlation coefficient (ICC). Fisher's exact test was used to assess differences between categorical measures.

Consistency between the 3 groups of raters was evaluated using Pearson's correlation coefficient and paired *t*-tests based on mean total NIH scores calculated for each case evaluated in each group of raters. For the comparison of intrarater scores, the mean total NIH score was calculated for each of the 12 cases for each group of evaluators. Using each group's mean total NIH score per case, pair-wise comparisons between groups were made using Pearson's correlation coefficient as a measure of agreement. Pair-wise differences in the mean total NIH scores per case were also calculated between groups and assessed using a paired *t*-test for each evaluation.

RESULTS

Forty study packets were distributed; 24 were returned for a 60% response rate. Two responders (8%) completed only a single evaluation. The demographic data of the 24 evaluators is summarized in Table 1. There was at least 1 evaluator (4%), and no more than 10 evaluators (42%) from each institution. Most (75%) considered themselves cGVHD "experts," which is reflected in the median years of experience (13, range: 0.5-32). Nearly half (42%) of evaluators reported previously receiving some type of formal training in the use of the instrument.

Interrater reliability statistics are summarized in Table 2. Overall mean kappa scores were highest for ulcers (0.46), followed by erythema (0.23), and lowest for lichenoid (0.15) and mucoceles (0.14; all *P*-values <.0001). Kappa scores were higher in OME compared with BMTE and BMT MLP in ulcers (0.85, 0.44, and 0.33, respectively) and erythema (0.40, 0.25, and 0.20, respectively). lichenoid kappa scores were highest in BMTMLP and similar in OME and BMTE (0.31, 0.16, and 0.11, respectively). Kappa scores for mucoceles were universally low. Rating of ulcers by OME was the only element that demonstrated *very good* agreement.

Overall intrarater reliability in all groups was high (overall ICC = 0.90, range: 0.88-0.95; Table 3). With respect to the component scores, there was much greater consistency in erythema and ulcers (overall ICC = 0.86 and 0.92, respectively) compared with lichenoid (0.72, range: 0.67-0.82) and mucoceles (0.73, range: 0.56-0.87).

For the comparison of intrarater scores, the mean total NIH score was calculated for each of the 12 cases for each group of evaluators at each evaluation (Table 4). Overall, there was high agreement in mean scores among the 3 groups of evaluators (r = 0.86-0.99). BMTE (group 2) and BMT MLP (group 3) on average, consistently had a higher mean total score

Table 1. Evaluator Demographics

	N (%)
Number of evaluators	24
Age (years), median (range)	44 (32, 59)*
Experience (years), median (range)	13 (0.5, 32)
Sex	
Female	8 (33)
Male	16 (67)
Institution	
University of North Carolina	3 (13)
University of Minnesota	1 (4)
Dana-Farber Cancer Institute/BWH	10 (42)
Fred Hutchinson Cancer Research Center	7 (29)
Northwestern University/Children's Memorial Hospital	2 (8)
Stanford University	1 (4)
Clinician type	
Medical staff ("BMT Expert")	16 (67)
Dental hygienist ("BMT MLP")	1 (4)
Oral medicine staff ("OM Expert")	4 (17)
Nursing staff ("BMT MLP")	3 (12)
Primary institutional responsibility	
Clinical	17 (71)
Research	4 (17)
Clinical and research	2 (8)
No response	1 (4)
Does rater consider him/herself a cGVHD expert?	
Yes	18 (75)
No	6 (25)
Has rater received formal training	
Yes	10 (42)
No	13 (54)
Unknown	1 (4)
Oral medicine specialist is part of clinical team	
Yes	16 (67)
No	8 (33)

OM indicates oral medicine; BMT MLP, bone marrow transplantation midlevel provider; cGVHD, chronic graft-versus host disease.
*Three evaluators did not report age.

than those of OME (group 1) (mean difference [95% confidence interval] = −0.81 [− 1.57, −0.05], −1.30 [− 2.07, −0.51], respectively, at the first evaluation; −1.09 [− 1.63, −0.54], −1.77 [− 2.50, −1.05], respectively, at the second evaluation), even though the magnitude of the difference was relatively small (Table 4).

Responses to the postevaluation questionnaire are summarized in Table 5. The most frequently reported areas of difficulty were (1) assessment of mucoceles (54%), (2) artifact from the camera flash (29%), (3) determining the percentage of involvement (17%). Although 75% of OME were *comfortable* with their abilities to score the cases, approximately 50% of BMTE and BMTMLP were *uncomfortable*. This is also reflected in the time required to complete the evaluations. The majority (58%) felt that their evaluations were *accurate*, with only 3 evaluators (13%, all BMTE) answering *inaccurate*. There was overwhelming consensus (84%) that formal training is required for accurate and effective use of the instrument.

DISCUSSION

This is the first study to evaluate inter- and intraobserver variability in the recently introduced NIH consensus criteria for scoring oral cGVHD. As the oral cavity is one of the most frequently affected sites [13,14], and easily examined clinically [18,19], we felt that this was an ideal target organ to examine variability parameters. Although the intent of this instrument is to provide an objective, reproducible score that accurately reflects the extent and severity of clinical disease, which generally correlates with symptoms [15,19], its functionality has yet to be demonstrated. Importantly, it was not our intent to evaluate the overall utility or validity of the scoring system, with respect to its ability to accurately measure disease severity and changes over time. In fact, to date, few studies using the NIH instruments have been published [11,20-22]. Elad et al. [23] recently reported only moderate correlation between NIH total scores and pain scores ($r = .45$, $P < .001$). Although they found strong correlations between total scores and erythema/ulceration scores, they also found that cases rated as "severe" based on erythema/ulceration versus lichenoid were significantly different, suggesting that such findings should not be classified at the same intensity level.

Overall interrater reliability ranged from *poor* to *fair*, with assessment of ulcers receiving the highest

Table 2. Summary of Interrater Reliability

Type	Overall Mean Kappa E1 (SE), E2 (SE)	OME Mean Kappa E1 (SE), E2 (SE)	BMTE Mean Kappa E1 (SE), E2 (SE)	BMT MLP Mean Kappa E1 (SE), E2 (SE)
No. of raters	23*	4	15	4
Type				
Erythema	0.23 (0.02), 0.23 (0.02)	0.45 (0.13), 0.35 (0.12)	0.26 (0.03), 0.24 (0.03)	0.14 (0.10), 0.26 (0.10)
Lichenoid	0.15 (0.02), 0.14 (0.02)	0.15 (0.01), 0.16 (0.14)	0.11 (0.03), 0.10 (0.03)	0.34 (0.09), 0.28 (0.08)
Mucoceles	0.17 (0.02), 0.12 (0.01)	0.08 (0.08), 0.02 (0.08)	0.16 (0.02), 0.13 (0.02)	0.11 (0.08), 0.02 (0.09)
Ulcers	0.45 (0.03), 0.47 (0.03)	0.80 (0.12), 0.89 (0.12)	0.42 (0.04), 0.45 (0.04)	0.33 (0.10), 0.32 (0.10)
Total*	0.29 (0.02), 0.28 (0.02)	0.56 (0.12), 0.56 (0.12)	0.29 (0.02). 0.28 (0.02)	0.14 (0.11), 0.14 (0.11)

E1 indicates evaluation #1; E2, evaluation #2; SE, standard error; $0 \leq \kappa \leq 0.20$ = *poor*, $0.21 \leq \kappa \leq 0.40$ = *fair*, $0.41 \leq \kappa \leq 0.60$ = *moderate*, $0.61 \leq \kappa \leq 0.80$ = *good*, and $0.81 \leq \kappa \leq 1.00$ = *very good;* OME, oral medicine expert; BMTE, bone marrow transplantation oncologist; BMT MLP, bone marrow transplantation midlevel provider.
*Rater 21 filled out (one initial, one follow-up); rater 21 filled out only the first evaluation, while rater 22 filled out only the second evaluation, resulting in 23 total raters at each evaluation, although 24 total participated.
*Total score was categorized as 0-5, 6-10, and 11-16.

Table 3. Intrarater Reliability

	Overall r, Mean ICC (Range)	OME r, Mean ICC (Range)	BMTE r, Mean ICC (Range)	BMT MLP r, Mean ICC (Range)
No. of raters	22*	4	14	4
Type				
Erythema	0.76, 0.86 (0.66, 1.00)	0.81, 0.91 (0.83, 1.00)	0.72, 0.83 (0.66, 0.93)	0.88, 0.94 (0.92, 1.00)
Lichenoid	0.67, 0.72 (0.31, 1.00)	0.65, 0.67 (0.39, 0.87)	0.65, 0.70 (0.31, 0.88)	0.72, 0.82 (0.67, 1.00)
Mucoceles	0.71, 0.73 (−0.33, 0.96)	0.80, 0.87 (0.77, 0.94)	0.71, 0.74 (0.18, 0.96)	0.56, 0.56 (−0.33, 0.94)
Ulcers	0.85, 0.92 (0.68, 1.00)	0.96, 0.98 (0.91, 1.00)	0.81, 0.89 (0.68, 1.00)	0.92, 0.96 (0.84, 1.00)
Total	0.83, 0.90 (0.61, 0.99)	0.90, 0.95 (0.93, 0.99)	0.79, 0.88 (0.61, 0.95)	0.88, 0.94 (0.91, 0.96)

r indicates Pearson's correlation coefficient; ICC, intraclass correlation coefficient; OME, oral medicine expert; BMTE, bone marrow transplantation oncologist; BMT MLP, bone marrow transplantation midlevel provider.
*Two BMTE raters only completed a single evaluation resulting in 22 rather than 24 evaluators.

kappa scores. Scoring of ulcers by the expert group demonstrated the highest scores in the range of *very good*; however, no other clinical features approached this level, regardless of the group evaluating. Ulcers can be easily differentiated from nonulcerated mucosa, the spatial assessment parameters (none, <20%, ≥20%) are simple to interpret, and there are only 3 options compared with 4 in the other features. This is in contrast with erythema, for example, where gradations may be more difficult to discern, and the spatial assessments are far more complex. Similar levels of modest concordance (ICC across sequential trials ranged from 0.57 to 0.70) between clinicians and experts using the NIH response criteria oral evaluation measures have been demonstrated by other investigators [24]. Of note, these investigators also observed that 83% of NIH response criteria clinician-assessed oral total scores (15 point instrument score range) were within 2 points of oral medicine experts' reference values.

Intrarater variability, in contrast, was far lower, with most overall mean kappa scores in the *very good* range. This is reflected in the finding that the majority (58%) felt *comfortable* with their ability to accurately score the cases. This suggests that with adequate training to ensure consistent use of the oral instrument (and by extension the other organ-specific instruments), accurate data can be collected in the context of cGVHD clinical trials. As the 60% response rate included representation from multiple centers, it is unclear to what extent there may have been significant differences between responders and nonresponders, which may have influenced study outcomes. The higher concordance rate among oral medicine specialists suggests that organ-specific experts should be involved in clinical trials at the level of training and calibration, and/or central data review.

The use of photographs allowed the participation of multiple clinicians at various centers throughout the United States. Although a similar approach of using clinical photographs to assess inter- and intraobserver variability in scoring has been previously employed, it is unclear how well interpretation of photographs reflects actual clinical observation, with respect to the ability to identify and discriminate specific findings [25,26]. Of note, >50% of evaluators reported difficulty assessing mucoceles, reflected in the *very poor* interrater variability for this feature (Table 2). This may have been explained by either the presence of flash artifact from the camera, or difficulty in discerning a raised, often colorless lesion in 2 dimensions, rather than evaluators' inability to clinically identify these lesions [27-29]. In fact, overall intrarater variability for mucoceles was *good*, demonstrating consistency regardless of any reported difficulties. Another reported complication was the ability to estimate percentages of involved mucosa, a critical component of evaluating lichenoid, erythematous, and ulcerative changes, suggesting that more specific guidelines might be necessary for optimal utilization of this instrument. Similar concerns with the ability

Table 4. Comparison of Total Scores between Groups

Correlation (r), (Mean Difference,* [95% CI])	Evaluation	OME	BMTE	BMTMLP	Mean (SE)
OME	E1	—	0.86 (−0.81, [−1.57, −0.05]†)	0.87 (−1.30, [−2.07, −0.51]†)	6.00 (0.60)
	E2	—	0.93 (−1.09, [−1.63,−0.54]†)	0.87 (−1.77, [−2.50, −1.05]†)	6.06 (0.62)
BMTE	E1	—	—	0.94 (−0.48, [−1.01, 0.05])	6.81 (0.71)
	E2	—	—	0.99 (−0.68, [−0.91, −0.46]†)	7.15 (0.69)
BMTMLP	E1	—	—	—	7.29 (0.74)
	E2	—	—	—	7.83 (0.71)

OME indicates oral medicine expert; BMTE, bone marrow transplantation oncologist; BMTMLP, bone marrow transplantation midlevel provider; CI, confidence interval.
*Mean differences are calculated as follows: Group 1-Group 2; Group 1-Group 3; Group 2-Group 3.
†Denotes the 95% confidence interval excludes zero; $P < 0.05$ for the paired *t*-test.

Table 5. Evaluators' Experiences Using the NIH Oral cGVHD Response Criteria Scoring System

	All	OME	BMTE	BMTMLP	*P*-value
N	24	4	16	4	
How long did this set of evaluations take to complete?					
≤30minutes	9 (37)	2 (50)	6 (38)	1 (25)	.87
>30minutes	12 (50)	2 (50)	7 (44)	3 (75)	
No response	3 (13)	0 (0)	3 (19)	0 (0)	
What part(s), if any, did you have trouble with?*					
Determining % involvement	4 (17)	2	2	0	
Spatial orientation	1 (4)	0	1	0	
Assessment of erythema	2 (8)	1	1	0	
Assessment of mucoceles	13 (54)	2	9	2	
Flash reflection	7 (29)	1	5	1	
Representation of ulcers	3 (13)	1	1	1	
Color representation/variation	2 (8)	1	0	1	
How comfortable were you with your ability to score these cases?					
Uncomfortable	9 (38)	0 (0)	7 (44)	2 (50)	.44
Neither/no response	3 (12)	1 (25)	2 (13)	0 (0)	
Comfortable	12 (50)	3 (75)	7 (44)	2 (50)	
How would you rate the accuracy of your evaluations?					
Inaccurate	3 (13)	0 (0)	3 (19)	0 (0)	>.99
Neither/No response	7 (29)	1 (25)	5 (31)	1 (25)	
Accurate	14 (58)	3 (75)	8 (50)	3 (75)	
Do you feel that formal training is required to accurately and effectively use this scoring system?					
Disagree	2 (8)	0 (0)	2 (13)	0 (0)	.83
Neither/no response	2 (8)	0 (0)	1 (6)	1 (25)	
Agree	20 (84)	4 (100)	13 (81)	3 (75)	

NIH indicates National Institutes of Health; cGVHD, chronic graft-versus-host disease; OME, oral medicine expert; BMTE, bone marrow transplantation oncologist; BMTMLP, bone marrow transplantation midlevel provider.
*Evaluators may have included more than one part that was considered problematic.

to reliably evaluate and score certain clinical features with respect to the skin cGVHD response criteria (ie, deep sclerosis) have been recently reported by others [22].

The instructions that were provided for the evaluators were comprehensive yet succinct, and did not include a specific training module or sample cases. This was intentional, so that there would be no potential bias if training compliance was variable. Of note, participants responded overwhelmingly (84%) that formal training should be required prior to application in the context of a clinical trial. Although we only evaluated the oral instrument, similar considerations for skin cGVHD evaluations would be expected given the instrument's various features that must be assessed (erythema, movable, and nonmovable sclerosis, ulceration, percentage of body surface area involved) [15,22]. Development of a coordinated training resource should be considered prior to commencing large-scale cGVHD clinical trials utilizing the new criteria, regardless of the level of experience and area of clinical expertise of those performing assessments. Of note, our data demonstrated that as long as clinicians are experienced in managing patients with cGVHD, their specific training or credentials have minimal, if any, impact on the ability to perform effective evaluations (Table 4).

Establishment of clinically meaningful and simple to use research instruments was a key outcome of the NIH conference [10,15]. Although the use of these instruments is certain to advance our understanding of cGVHD, studies such as the present are critically important in defining their strengths and weaknesses so that they can be further refined and modified for optimal utilization. Equally as important, studies are needed to assess the instruments validity and clinical significance in the context of interventional clinical trials. Such initiatives can only be achieved by multiinstitutional collaborations, with the common goal of minimizing the morbidity of cGVHD.

AKNOWLEDGMENTS

The authors would like to thank all of the clinicians that participated in this study for their time and care in completing the evaluations. They are particularly grateful for the contributions of Mr. Terrence Hanscom and Ms. Marianne O'Shea for their help with preparing the study packets and data coordination and processing.

Financial disclosure: C. Cutler is supported by the Stem Cell Cyclists of the Pan-Mass Challenge.

REFERENCES

1. Gross TG, Egeler RM, Smith FO. Pediatric hematopoietic stem cell transplantation. *Hematol Oncol Clin North Am*. 2001;15:795-808.
2. Bhushan V, Collins RH Jr. Chronic graft-vs-host disease. *JAMA*. 2003;290:2599-2603.
3. Lee SJ, Vogelsang G, Flowers ME. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2003;9:215-233.

4. Leger CS, Nevill TJ. Hematopoietic stem cell transplantation: a primer for the primary care physician. *CMAJ*. 2004;170:1569-1577.
5. Higman MA, Vogelsang GB. Chronic graft versus host disease. *Br J Haematol*. 2004;125:435-454.
6. Syrjala KL, Chapko MK, Vitaliano PP, Cummings C, Sullivan KM. Recovery after allogeneic marrow transplantation: prospective study of predictors of long-term physical and psychosocial functioning. *Bone Marrow Transplant*. 1993;11:319-327.
7. Chiodi S, Spinelli S, Ravera G, et al. Quality of life in 244 recipients of allogeneic bone marrow transplantation. *Br J Haematol*. 2000;110:614-619.
8. Antin JH. Clinical practice. Long-term care after hematopoietic-cell transplantation in adults. *N Engl J Med*. 2002;347:36-42.
9. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8:444-452.
10. Pavletic S, Vogelsang G. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: preface to the series. *Biol Blood Marrow Transplant*. 2005;11:943-944.
11. Pavletic SZ, Lee SJ, Socie G, Vogelsang G. Chronic graft-versus-host disease: implications of the National Institutes of Health consensus development project on criteria for clinical trials. *Bone Marrow Transplant*. 2006;38:645-651.
12. Ratanatharathorn V, Ayash L, Lazarus HM, Fu J, Uberti JP. Chronic graft-versus-host disease: clinical manifestation and therapy. *Bone Marrow Transplant*. 2001;28:121-129.
13. Flowers ME, Parker PM, Johnston LJ, et al. Comparison of chronic graft-versus-host disease after transplantation of peripheral blood stem cells versus bone marrow in allogeneic recipients: long-term follow-up of a randomized trial. *Blood*. 2002; 100:415-419.
14. Lee SJ, Klein JP, Barrett AJ, et al. Severity of chronic graft-versus-host disease: association with treatment-related mortality and relapse. *Blood*. 2002;100:406-414.
15. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: IV. Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2006;12:252-266.
16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
18. Imanguli MM, Alevizos I, Brown R, Pavletic SZ, Atkinson JC. Oral graft-versus-host disease. *Oral Dis*. 2008;14:396-412.
19. Treister NS, Cook EF Jr., Antin J, Lee SJ, Soiffer R, Woo SB. Clinical evaluation of oral chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2008;14:110-115.
20. Sari I, Altuntas F, Kocyigit I, et al. The effect of budesonide mouthwash on oral chronic graft versus host disease. *Am J Hematol*. 2007;82:349-356.
21. Cho BS, Min CK, Eom KS, et al. Feasibility of NIH consensus criteria for chronic graft-versus-host disease. *Leukemia*. 2009;23: 78-84.
22. Jacobsohn DA, Rademaker A, Kaup M, Vogelsang GB. Skin response using NIH consensus criteria vs Hopkins scale in a phase II study for steroid-refractory chronic GVHD. *Bone Marrow Transplant*. 2009.
23. Elad S, Zeevi I, Shapira M. Validation of the NIH Scale for Oral GVHD. Oral Complications of Emerging Cancer Therapies Conference. Bethesda, MD, April 14–15, 2009.
24. Mitchell S, Jacobsohn D, Thormann K, et al. Feasibility and reproducibility of the NIH consensus criteria to evaluate response in Chronic Graft Versus Host Disease (cGvHD), American Society of Hematology Annual Meeting, Orlando, FL, 2006: Vol. 108.
25. Pandolfino JE, Vakil NB, Kahrilas PJ. Comparison of inter- and intraobserver consistency for grading of esophagitis by expert and trainee endoscopists. *Gastrointest Endosc*. 2002;56: 639-643.
26. Piboonniyom SO, Treister N, Pitiphat W, Woo SB. Scoring system for monitoring oral lichenoid lesions: a preliminary study. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 2005; 99:696-703.
27. Garcia FVMJ, Pascual-Lopez M, Elices M, Dauden E, Garcia-Diez A, Fraga J. Superficial mucoceles and lichenoid graft versus host disease: report of three cases. *Acta Derm Venereol*. 2002;82: 453-455.
28. Demarosi F, Lodi G, Carrassi A, Sardella A. Superficial oral mucoceles: description of two cases in patients with graft-versus-host disease. *J Otolaryngol*. 2007;36:E76-E78.
29. Balasubramaniam R, Alawi F, DeRossi S. Superficial mucoceles in chronic graft-versus-host disease: a case report and review of the literature. *Gen Dent*. 2009;57:82-88.